DOCUMENT RESUME

ED 362 537 TM 020 555

AUTHOR Hambleton, Ronald K.; Kanjee, Anil
TITLE Enhancing the Validity of Cross-Cultural Studies: Improvements in Instrument Translation Methods.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Apr 93
NOTE 20p.; Paper presented at the Annual Meetings of the American Educational Research Association (Atlanta, GA, April 12-16, 1993) and the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Attitude Measures; *Cross Cultural Studies; Cultural Differences; *Culture Fair Tests; Elementary Secondary Education; *Psychological Testing; Test Construction; Test Interpretation; Test Items; *Test Validity; *Translation
IDENTIFIERS Speededness (Tests); *Test Equivalence

ABSTRACT

Translating psychological tests and attitude scales from one language and culture for use in others has been common in the history of educational and psychological assessment. Despite the long history, methods for conducting test and attitude scale translations and establishing equivalence of the multiple versions of an instrument are not well-known or routinely used in practice. The primary purpose of this paper is to focus attention on three major sources of invalidity in test translation work and to suggest solutions whenever possible. The first is cultural differences associated with administrations, item formats, and speededness. The second is technical factors associated with the instrument itself, the selection and training of translators, the translation process, and the judgmental and empirical designs that can be used in translating instruments. The third source of invalidity is factors affecting the interpretations of achievement test results such as similarity of curricula, equivalence of motivational levels, and sociopolitical considerations. (Contains 21 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 537

Enhancing the Validity of Cross-Cultural Studies: Improvements in Instrument Translation Methods

Ronald K. Hambleton and Anil Kanjee
University of Massachusetts at Amherst

Abstract

Translating psychological tests and attitude scales from one language and culture for use in others has been common in the history of educational and psychological assessment. Despite the long history, methods for conducting test and attitude scale translations and establishing equivalence of the multiple versions of an instrument are not well-known or routinely used in practice. The primary purpose of this paper is to focus attention on three major sources of invalidity in test translation work and to suggest solutions whenever possible: The first is cultural differences associated with administrations, item formats, and speededness. The second is technical factors associated with the instrument itself, the selection and training of translators, the translation process, and the judgmental and empirical designs that can be used in translating instruments. The third source of invalidity is factors affecting the interpretations of achievement test results such as similarity of curricula, equivalence of motivational levels, and socio-political considerations.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RONALD K. HAMBLETON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM020555

Enhancing the Validity of Cross-Cultural Studies: Improvements in Instrument Translation Methods[1,2,3,4]

Ronald K. Hambleton and Anil Kanjee
University of Massachusetts at Amherst

For as long as psychological tests and attitude scales have existed, researchers have been interested in translating them. One reason is that it is often cheaper and faster to translate an instrument than it is to develop a new instrument for a second language group. Sometimes, too, the technical expertise does not exist in the second language group to actually construct the needed instrument.

A second reason is that translated tests and scales allow cross-national, cross-language, and/or cross-ethnic comparative studies to take place. Such studies have become particularly popular in recent years as many countries strive to set world-class educational standards or simply to look at their own educational progress in relation to other countries. For example, over 60 countries will participate in the Third International Mathematics and Science Study (TIMSS) being conducted in 1994 and 1998 by the IEA.

Finally, instruments are translated to enhance fairness in assessment by enabling persons to take tests and psychological scales in their preferred languages. For example, high school students in Israel can take their college

---

[1]Paper presented at the meetings of AERA and NCME, Atlanta, 1993.

[2]Laboratory of Psychometric and Evaluative Research Report No. 255. Amherst, MA: University of Massachusetts, School of Education.

[3]To appear in T. Husen and T. N. Postlethwaite (Eds.). (1994). International Encyclopedia of Education (2nd ed.). Oxford, UK: Pergamon Press.

[4]Partial support for the preparation of this paper was provided by the National Center for Education Statistics (NCES), Department of Education. The opinions expressed, however, are those of the authors and not necessarily the NCES.

admission exams in one of six languages. Thus, the bias in exam scores associated with examinees being examined in their second or third best language is removed and exam score validity is enhanced. While an effort is made to properly translate the exams into six languages, what translation problems remain are judged to be less serious from a validity perspective than the unfairness that would result from requiring all exam candidates to take their exams in Hebrew.

While the reasons for translating tests and attitude scales are clear, the methods for doing the translations and establishing the equivalence of scores from the two versions of the instrument are not (Hambleton, 1993). Some cross-cultural researchers have even speculated that a high percentage of research in their field is flawed to the point of being invalid because of the use of poorly translated instruments. The purposes of this paper are to review the sources of invalidity associated with translated tests and scales and to suggest solutions whenever possible. The sources of invalidity can be organized into three broad categories: cultural/language differences, technical issues and methods, and interpretation of results. For the purposes of this paper, the terms - tests, scales, instruments, and assessment measures - will be used interchangeably and translation work will be used in the broad sense to include adaptation work which may sometimes be needed.

1. Cultural/Language Differences Affecting Scores

The assessment and interpretation of cross-cultural results cannot be viewed in the narrow context of just the translation or adaptation of instruments. Rather, this process should be considered for all parts of the assessment process including the administration of instruments, item formats used, and the effect of speed on examinee performance. These three factors will be considered next.

Test Administration. Communication problems between examiner and examinees can prove to be a serious threat to the validity of results. Van de Vijver and Poortinga (1991) noted the communication failure experienced by Greenfield (1966, 1979) when assessing the principle of conservation among the Wolof people. Greenfield presented subjects with two differently shaped beakers, one tall and one broad, that contained equal amounts of water. Subjects were asked to identify the beaker which contained more water. Greenfield found that subjects frequently responded with answers that failed to show an ability to conserve, that is, many students said that the tall beaker contained more water. However, Irvine (1978) later found that in the Wolof language, "more" referred to both quantity and level, and demonstrated (by posing the question differently) that the Wolof did in fact have mastery of the principle of conservation.

One way to circumvent problems between examiner and examinees is to ensure that the instructions on the test itself are clear and self-explanatory, with minimal reliance on verbal communication (van de Vijver & Poortinga, 1991). Special problems can be expected with directions for rating scales used in attitude measurement too since they are not common in many countries. Also, test administrators should (1) be drawn from the target communities, (2) be familiar with the culture, language, and dialects, (3) have adequate test administration skills and experience, and (4) possess some measurement expertise. Additionally, consistency in test administration across different groups can be improved by providing (basic) training to all test administrators. Training sessions should be pre-planned as part of the test development process, stressing clear, unambiguous communication, the importance of following instructions, strictly following time limits, the influence of test administrators on reliability and validity, etc.

Test Format. Differential familiarity with particular item formats presents another source of invalidity of test results in cross-cultural studies. For example, in the U.S., selected response questions, especially multiple-choice questions, have been used extensively in assessment. In cross-cultural studies, it cannot be assumed that all students are as familiar with multiple choice items as U.S. students. Nationalities that follow the British system of education place greater emphasis on essays and short answer questions, as opposed to multiple-choice items. Thus, students from these countries are placed at a possible disadvantage as compared to their U.S. counterparts. Of course, when constructed response formats are emphasized or serve as the dominant mode of assessment, persons with more experience with selected response formats such as multiple-choice item formats will be placed at a disadvantage. Sometimes a balance of item formats may be the best solution to insure fairness and reduce sources of invalidity in the assessment process.

Another solution is to include only those formats with which all groups being assessed are experienced. Whenever it can be assured that examinees are not placed at a disadvantage, and when all variables of interest can still be measured, multiple-choice items or simple rating scales should be preferred. The major advantage is that multiple-choice items or simple rating scales can be objectively scored. Thus, complications in scoring associated with open-ended responses are avoided. This is especially relevant in cross-cultural studies where it may be more difficult to translate the scoring rubrics than the test items! Also, practice items can easily be included to enable examinees to familiarize themselves with the "different" formats. In addition, extensive, unambiguous instructions including examples and exercises help to reduce differential familiarity (van de Vijver & Poortinga, 1992).

A common assumption is that examinees easily grasp the meaning of pictorial stimuli. However, pictorial stimuli are subject to bias like any other stimuli, as perception is strongly influenced by previous experience (Lonner, 1990).

Speed. It is often assumed that examinees will work fast on "speeded" tests (van de Vijver & Poortinga, 1991). However, in a study comparing Dutch and other ethnic students in Holland, Van Leest and Bleichrodt (1990) found that the speed factor increased ethnic bias. Schmidtt and Crone (1991), in a differential item functioning study comparing Whites and Hispanic Americans, also found speededness to be a factor negatively affecting Hispanic examinees' performance. In a study that compared cognitive gender differences between German and American examinees, Ellis and Weiner (1990) found no gender differences, in both American and German examinees, when speed of administration was minimized. The best solution would seem to be to minimize test speededness as a factor in cognitive test performance and attitude scales unless there is a good reason for including it.

2. Technical Issues and Methods

In this section, five technical factors which can influence the validity of translated instruments are considered: the instrument itself, selection and training of translators, the process of translation, judgmental designs for translating instruments, and empirical designs for establishing equivalence.

The Instrument Itself. If a researcher knows that he/she will be using a test or attitude scale in a different language or culture, it is advantageous to take this into account at the outset of the instrument development process. Failure to do so can introduce problems later in the translation process which will reduce the validity of the translated

instrument. Choice of item formats, stimulus material for the instrument, vocabulary, sentence structure, and other aspects which might be difficult to translate well can all be taken into account in the initial instrument development to minimize problems later in translation. For example, passages about the game of baseball which would be unfamiliar in many cultures could be rejected in favor of passages about walking through a park or other activities that would have meaning across many language and cultural groups. Units of measurement should be avoided too since they vary from one nationality to another.

With attitude scales especially, care must be taken to choose situations, vocabulary, and expressions that will translate easily across language groups and cultures. Very often these scales contain everyday expressions which enhance their meaningfulness in one language but make translations difficult. For example, an expression such as "What goes around, comes around" could be difficult to translate.

Selection and Training of Translators. The importance of obtaining the services of competent translators should be obvious. Too often though, researchers have tried to go through the translation process with a single translator selected because he/she happened to be available. Competent translation work cannot be assumed. Also, the use of a single translator, competent or not, does not permit highly valuable discussions of independent translations across a group of competent translators.

But translators should be more than persons familiar and competent with the languages involved in the translation. They should know the cultures very well especially the target culture (i.e., the culture of the language in which the instrument is being translated). Knowledge of the cultures involved especially the target culture is often essential for an effective translation.

Also, subject matter knowledge in the translation of achievement tests is essential. The nuances and subtleties of a subject area will be lost on an translator unfamiliar with the subject matter. In one project we heard about recently, a translator translated the term "item pools" used in test development work to "item oceans" in Japanese. Too often, translators without technical knowledge will resort to literal translations which are often problematic to target language examinees and threaten test validity.

Finally, test translators would benefit from some training in test and attitude scale construction. For example, test translators need to know that when doing translations they should not create clang associations that might lead test-wise examinees to the correct answers, or translate distractors in multiple-choice items unknowingly so that they have the same meaning. A test translator without any knowledge of the principles of test and scale construction could easily make test material more or less difficult unknowingly, and correspondingly, lower the validity of the instrument in the target population.

Process of Translation. The problem of dialects within a language can become a threat to the validity of translated tests. Which dialect is of interest, or is the goal to produce a translation that could apply across dialects within a language? This problem should be resolved, used in the selection of translators, and addressed in the training of translators.

Frequency counts of words can be invaluable in producing valid translations. In general, it is best to translate words and expressions with words and expressions with approximately the same frequencies in the two languages. One additional problem is that these frequency lists of words and expressions are not always available. This again is the reason for preferring

translators who are familiar with both of the cultures and not just the languages.

"Decentering" is sometimes used in translating instruments. It may be that some words and expressions do not have equivalent words and expressions in the target language. It is even possible that the words and expressions do not exist in the target language. Decentering involves making revisions to the source language instrument so that equivalent material can be used in both the source and target language versions of the instrument. Such a strategy is possible when the source language instrument is under development at the same time as the target language version.

Judgmental Designs for Translating Instruments. The two most popular judgmental designs are forward translations and backward translations. With forward translations, a single translator or preferably a group of translators, translate the test or attitude scale from the source language to the target language. Then, the equivalence of the two versions of the instrument is judged by another group of translators. Revisions can be made to the target version of the instrument to correct problems identified by the translators. Sometimes the validity of the judgments about the equivalence of the two versions is enhanced by also having examinees provide translators or a group of judges with their interpretations of the material on the tests and questionnaires. The basic design is weak however because of the high level of inferencing that must be done by the translators or the judges about the equivalence of the two versions of the instrument.

The back-translation design is the best known and most popular of the judgmental designs. In one variation, a group of translators translates the instrument from the source language to the target language. A second group of translators takes the translated instrument (in the target language) and

translates it back to the source language. Then, the original version of the instrument and the back-translated version are compared and judgments are made about their equivalence. To the extent that the two versions of the instrument in the source language look similar, support is available for the equivalence of the source and target versions of the instrument. The back-translation design can be considered as a general check on translation quality that can detect at least some of the problems associated with poor translations or adaptations. It has been used successfully in many projects as a first step in assessing the quality of a translation.

Though the back-translation design is to be recommended for use in many projects, it would rarely provide a sufficient amount of evidence to support the use of a translated instrument in practice. (This design may suffice in small scale minor cross-cultural studies.) Evidence of instrument equivalence provided by a back-translation design is only one of many types of evidence that should be compiled in a translation study. One of the main shortcomings is that the comparison of instruments is carried out in the source language. It is quite possible that the translation could be poor while the evidence on the comparability of the original instrument and the back-translated instrument would suggest otherwise. This might happen if the translators used a shared set of translation rules that insured that the back-translated instrument looked like the original instrument. A second shortcoming is that the translation could be poor because it retained inappropriate aspects of the source language instrument such as the same grammatical structure and spelling. Such errors facilitate back-translations but they mask serious shortcomings in the target version of the instrument. Finally, this and other judgmental designs can be faulted because samples of the intended populations for the instruments never actually take the instruments under test-like

conditions (or, for that matter, any other conditions). There is very little evidence to support the position that translators or other judges are capable of predicting the equivalence of versions of an instrument from a review, however carefully it may be done. In fact, most of the available evidence suggests that judges are not very successful at predicting test items that function differentially in two or more groups.

Empirical Designs for Establishing Equivalence. The empirical designs are of two types: those that use bilingual participants and those that use monolingual participants in the source and target languages. Designs that use bilinguals are often difficult to carry out because of the shortage of bilinguals who are equally proficient in both languages. And, when the samples taking each version of the instrument are not carefully matched on ability, only simple and not very informative statistical analyses can be carried out. For example, the relative order of item difficulty in the two versions of the instrument might be checked.

Even when an appropriate sample of bilingual participants can be found to take one or both versions of the instrument, problems remain. For one, evidence of equivalence of two versions of an instrument (such as similar item statistics, score distributions, and factor structures) in a bilingual sample of persons may not generalize to the monolingual persons in each population. For example, in a study by Hulin, Drasgow, and Komocar (1982) with the Job Descriptive Index, they found that with a bilingual sample of participants, only 4% of the items in the attitude scale were identified as poorly translated. The result jumped to 30% of the items when monolingual samples of participants from the source and target language populations were used.

A better empirical design would involve monolinguals taking the source language version of the instrument and a second sample of monolinguals taking

the target language version of the instrument. An assumption of equal ability distributions across the two groups is not usually tenable but it is still possible to compare item statistics if the analyses are carried out within an item response theory framework (Ellis, 1989; Hambleton, Swaminathan, & Rogers, 1991), or other statistical frameworks which are not based on an assumption of equal ability distributions. The advantages of this design are that samples of the source and target populations are used in the analyses and therefore findings about the equivalence of the two versions of the instrument are generalizable to the populations of interest. These studies are carried out like item bias studies (Hambleton, Swaminathan, & Rogers, 1991). Comparisons of the item statistics in the two versions of the instrument are made controlling for any ability differences in the two groups. Items showing differences are identified and carefully studied to determine the possible explanations for the differences. A poor translation is one likely explanation.

3. Interpretation of Results

In large-scale cross-cultural studies, the purpose of the instrumentation is to provide a basis for making comparisons between various cultural/language groups, so as to understand the differences and similarities that exist (Hambleton & Bollwark, 1991). Sometimes cognitive variables are of interest and other times the focus may be on the assessment of personality variables or general information. Results should be used for seeking ways of comparing groups and understanding the differences. Cross-cultural studies should not be used to support arguments about the superiority or exceptionality of nations as if it were some sort of a horse race (Westbury, 1992). In this context, to gain a better understanding when interpreting scores, other relevant factors external to the tests or assessment measures

and specific to a nationality should also be considered. Curricula, educational policies, wealth, standard of living, cultural values, etc. may be essential for properly interpreting scores across cultural/language and/or national groups. Next, several of the factors which should be considered in interpreting achievement test results across groups will be presented.

Similarity of Curricula. To the extent that differences in curricula exist, any achievement comparisons between different cultures will be tenuous if these curricula differences are not taken into account. Westbury (1992) notes that the results of the Second International Mathematics Study (SIMS) indicate that U.S. students performed poorly in every grade and in every aspect of mathematics tested. When comparing performance of Japanese and U.S. students, major curricular differences between the two countries were noted. However, when the curricula of the two countries were similar, Westbury found no essential differences between the performance of U.S. and Japanese students.

In another example, Song and Ginsburg (1988) compared the early learning behavior of Korean and U.S. children and found that superior achievement in mathematics in Korean children could be attributed to the dual number systems taught in Korean (Chinese and English systems). In addition to other cultural factors like parental involvement and teacher-student relationship, they also found that the amount of time devoted to mathematics in Korea was greater than that in the U.S.

Under these different conditions, it is not unusual to expect differences in performance. Overlooking the specific and unique national characteristics that affect the test scores can have serious consequences on the interpretation of results. Perhaps these omissions help in explaining some of the recent IAEP and IEA results.

Student Motivation. Wainer (1993) questioned whether demonstrated proficiency as measured by tests can be separated from motivation. He noted that in the recent International Assessment of Educational Progress study (Lapointe, Mead, & Askew, 1992), all the (randomly) selected Korean students were made aware of the great honor of being chosen to represent their school and country, and thus had a responsibility to perform at their best. For American students, on the other hand, participation on this international comparative study was just another activity.

Also, van de Vijver & Poortinga (1991) noted that it cannot be assumed that examinees will always try to achieve a high score. For example, for Black South African students, the aim in exams is to achieve the minimum score required to pass any exam. This is because the imposed state education system is perceived by many examinees to be detrimental to Blacks, and thus students only aspire for the minimum required of them. In this context, it is not unusual to expect different levels of performance, which may have very little to do with ability.

Socio-Political Factors. The meaning and interpretation of test scores can also differ even though the scores may be equivalent. For example, consider comparing test scores between students from developed and developing nations, or industrialized and mainly rural societies. In this context, performance of students may not be related to ability at all. Rather, it may be a reflection of the lack of access to adequate resources, or the different quality of educational services available.

The point is that, for any meaningful interpretations, the different social, political and economic realities facing nationalities, as well as the relevance of educational opportunities in the light of these realities must be considered (Olmedo, 1981). Thus it is important for test developers to be

aware of those specific cultural issues that might impact on test scores. Test developers and translators familiar with the target nationality play a crucial role in this regard.

4. Emerging Issues and Research

Currently, there do not exist technical standards for conducting translation studies that have the support and approval of international psychology organizations whose members do cross-cultural testing and research. This is an important point because of the ever expanding interest in translating instruments. Fortunately such a set of standards are now being developed. The International Test Commission has organized an international committee of psychologists from the IEA, the European Association of Psychological Assessment, the International Association of Cross-Cultural Psychology, the International Association of Applied Psychology, and the International Union of Psychological Science to prepare a validated set of technical standards. The work of this 13-person committee which represents six international organizations should be completed by the spring of 1994. Over 40 psychologists throughout the world have agreed to participate in the field test of the technical standards. The availability of technical standards for translating instruments should facilitate the proper translation of tests and attitude scales and the compilation of evidence to support the intended uses of these instruments.

Currently, not only are more tests and attitude scales being translated than ever before, but the tests and scales are being put to important uses by national governments such as establishing world-class performance standards. The validity of scores, therefore, from translated tests and scales must be clearly established. The consequence is that more sophisticated methodology is being used to establish equivalence. Item response theory models (Hulin,

1987) are being used to identify poorly translated items and to place scores from different translations of a test or scale onto the same reporting scale. The specific details associated with model selection, test score linking designs, and identification of problematic items in the translation process still remain to be worked out. In principle, the solutions are known but considerably more experience with translated instruments possibly involving limited sample sizes is needed.

Structural equation models (Joreskog & Sorbom, 1993) are being used too in the study of factor structures underlying tests and scales in multiple language groups. Such analyses are central in assessing the equivalence of instruments across cultural/language groups. Clearly, methodological advances are needed to insure that the equivalence of translated tests and scales can be adequately determined. One special problem might be the study of equivalence in factor structures across many language groups and with modest sample sizes in each.

5. Summary

To enhance the meaning of any cross-cultural research, it is important for researchers to carefully choose test administrators, use appropriate item formats, and control for the speed effect. In addition, translators who are familiar with the target group and their language, who know the content of the instrument, and who have received some training in instrument development, are the most capable persons for getting the translation job done well. Appropriately chosen judgmental designs (such as backward translations) and empirical designs and analyses (such as comparisons of results from monolingual examinees taking the instrument in their own language) can provide invaluable data bearing on the question of instrument equivalence across groups. With regard to interpretation of scores, those specific background

variables that impact on performance should be carefully considered. In this regard, differing curricula, levels of motivation and socio-political factors are especially important with achievement tests. Also, comparisons should not only be undertaken with emphasis on the differences. Similarities between nationalities can also provide useful and relevant information.

References

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. Journal of Applied Psychology, 74, 912-921.

Ellis, B. B., & Weiner, S. P. (1990). A study of the gender differences in two countries: Implications for future research. In N. Bleichrodt & P. J. D. Drenth (Eds.), Contemporary issues in cross-cultural psychology. Amsterdam, The Netherlands: Swets & Zeitlinger.

Greenfield, P. M. (1966). On culture and conservation. In J. S. Bruner, R. R. Olver & P. M. Greenfield (Eds.), Studies in cognitive growth (pp. 225-256). New York: Wiley.

Greenfield, P. M. (1979). Response to Wolof "magical thinking." Journal of Cross-Cultural Psychology, 10, 251-256.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 54-65.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: Technical issues and methods. Bulletin of the International Test Commission, 18, 3-32.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. Journal of Cross-Cultural Psychology, 18, 115-142.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translation. Journal of Applied Psychology, 67, 818-825.

Irvine, J. T. (1978). Wolof "magical thinking": Culture and conservation revisited. Journal of Cross-Cultural Psychology, 9, 300-310.

Jöreskog, K. G., & Sörbom, D. (1986). LISREL8: Structural equation modeling with the SIMPLIS command language. Hillsdale, NJ: Erlbaum.

Lapointe, A. E., Mead, N. A., & Askew, J. M. (1992). Learning mathematics (Report No. 22-CAEP-01). Princeton, NJ: Educational Testing Service.

Lonner, W. J. (1990) An overview of cross-cultural testing and assessment. In R. W. Brislin (Ed.), Applied cross-cultural psychology (Volume 14) (pp. 56-76). Newbury Park, CA: Sage Publications.

Olmedo, E. L. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.

Schmitt, A. P., & Crone, C. R. (1991). Alternative mathematical aptitude item types: DIF issues (Research Report 91-42). Princeton, NJ: Educational Testing Service.

Song, M. J., & Ginsburg, H. P. (1988). The effect of the Korean number system on young children's counting: A natural experiment in numerical bilingualism. International Journal of Psychology, 23, 319-332.

van Leest, P. F., & Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. In N. Bleichrodt & P. J. D. Drenth (Eds.), Contemporary issues in cross-cultural psychology. Amsterdam, The Netherlands: Swets & Zeitlinger.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing (pp. 277-308). Boston, MA: Kluwer Academic Publishers.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? European Journal of Psychological Assessment, 8, 17-24.

Wainer, H. (1993). Measurement problems. Journal of Educational Measurement, 30, 1-21.

Westbury, I. (1992). Comparing American and Japanese achievement: Is the United States really a low achiever? Educational Researcher, 21, 18-24.